

# On Randomized Distributed Coordinate Descent with Quantized Updates

Mostafa El Gamal  
Dept. of Electrical and  
Computer Engineering  
Worcester Polytechnic Institute  
Worcester, Massachusetts  
Email: melgamal@wpi.edu

Lifeng Lai  
Dept. of Electrical and  
Computer Engineering  
University of California, Davis  
Davis, California  
Email: llai@ucdavis.edu

**Abstract**—In this paper, we study the randomized distributed coordinate descent algorithm with quantized updates. In the literature, the iteration complexity of the randomized distributed coordinate descent algorithm has been characterized under the assumption that machines can exchange updates with an infinite precision. We consider a practical scenario in which the messages exchange occurs over channels with finite capacity, and hence the updates have to be quantized. We derive sufficient conditions on the quantization error such that the algorithm with quantized update still converge. We further verify our theoretical results by running an experiment, where we apply the algorithm with quantized updates to solve a linear regression problem.

**Index Terms**—Distributed optimization, coordinate descent, quantization error, convergence rate, quantization-free algorithm.

## I. INTRODUCTION

Developing algorithms to solve distributed optimization problems has attracted significant interests in recent years. Distributed optimization problems naturally arise in various scenarios. For example, in machine learning problems, the training dataset might be too large to be stored in a single machine. Another example is when data is collected (and hence is naturally located) at multiple locations. Distributed algorithms are also useful to harness parallel processing capabilities of multiple machines.

In distributed optimization, it is essential for machines involved to exchange messages. As communication links between machines have limited capacity and have significantly longer delay, many recent papers focus on developing algorithms that are communication efficient. In [1], an algorithm was proposed to reduce the amount of necessary communication by using the local computation in a primal-dual setting. Another communication efficient algorithm for empirical risk minimization was introduced in [2]. ADMM was considered in [3], [4], [5] to handle the communication bottleneck.

Most of the existing studies analyze how many rounds of communications are required for the convergence of the developed algorithms. In each communication round, it is typically assumed that machines can exchange messages with an infinite precision. However, in practice, these data exchanges occur over physical channels that have limited capacity. As a result, machines cannot exchange messages with an infinite

precision and need to quantize messages before sending them to other machines. A natural question to ask is whether these distributed algorithms will still converge if the exchanged messages are quantized. If these algorithms still converge, one can further ask what are the effects of the quantization on the converge speed.

In this paper, we answer these questions for a particular optimization algorithm, namely randomized coordinate descent [6]. This algorithm is easily implementable to solve distributed optimization problems since each machine can compute a single coordinate of the gradient. In each iteration of the randomized coordinate descent, the algorithm takes a step in the direction of a randomly chosen coordinate in order to decrease the function value. This is done by computing the partial derivatives, which is much cheaper computationally than taking a full gradient step. The iteration complexities of the randomized coordinate descent algorithms are analyzed in [7], [8] under a very general setup. In [9], a hybrid coordinate descent method (Hydra) was presented to speed up the coordinate descent algorithm. Asynchronous parallel processing was analyzed in [10] for a number of optimization algorithms including the randomized coordinate descent.

We answer the above questions by first modifying a distributed version of the coordinate descent algorithm to fit the paradigm of capacity limited communication. We then determine sufficient conditions on the quantization error such that the algorithm converges to the optimal solution. In particular, we apply our algorithm to an unconstrained minimization problem of a function  $f$  that is  $L$ -smooth and  $m$ -strongly convex. We show that for an accuracy level  $\epsilon$  and a confidence level  $\rho$ , our algorithm converges to the optimal solution if the quantization error  $\Delta$  is upper bounded by a function of  $\epsilon$ ,  $\rho$ ,  $L$ ,  $m$ , and  $d$ , where  $d$  is the number of features. We verify the results by running an experiment, where we apply our algorithm to solve a linear regression problem.

The rest of the paper is organized as follows. We give a formal statement of the problem in Section II. In Section III we introduce our algorithm. We analyze the convergence rate of our algorithm, and we derive sufficient conditions on the quantization error in Section IV. We verify our results by running an experiment in Section V. Finally, we conclude the

paper in Section VI.

## II. PROBLEM FORMULATION

We consider an unconstrained convex minimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}), \quad (1)$$

where  $\mathbf{x} = \{x_1, x_2, \dots, x_d\}$ , and  $f(\mathbf{x})$  is an  $L$ -smooth and  $m$ -strongly convex function, such that for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , we have that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad (2)$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq m\|\mathbf{x} - \mathbf{y}\|^2, \quad (3)$$

where  $L$  is the Lipschitz constant and  $m$  is the strong convexity parameter. The condition number of  $f$  is defined as  $g = L/m$ . As a result of the strong convexity, the function  $f(\mathbf{x})$  has a unique minimum at  $\mathbf{x}^*$ .

In the distributed coordinate descent algorithm, the data examples related to the problem are distributed over  $d$  nodes such that each node can calculate one coordinate of the gradient  $\nabla f(\mathbf{x})$  as explained in Section 6 of [9]. The algorithm we study in this paper is the randomized coordinate descent, in which at each iteration a coordinate is randomly selected to be updated. There are different ways to randomly select the coordinate. In this paper, we focus on the case in which the coordinates are selected with a uniform distribution. The channels connecting machines are capacity limited with a quantization resolution of  $\Delta$ , which means that machine  $i$  can only send a quantized version  $Q\left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right)$  of its update  $\frac{\partial f(\mathbf{x})}{\partial x_i}$ , such that

$$Q\left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right) = y\Delta, \text{ if } (y - \frac{1}{2})\Delta \leq \frac{\partial f(\mathbf{x})}{\partial x_i} < (y + \frac{1}{2})\Delta, \quad (4)$$

in which  $Q(\cdot)$  is the quantization operator. Let  $[\nabla f(\mathbf{x})]_i \in \mathbb{R}^d$  denote a vector that has only one nonzero element at position  $i$  that is equal to  $\frac{\partial f(\mathbf{x})}{\partial x_i}$ . By applying the quantization operator to the nonzero element of the vector  $[\nabla f(\mathbf{x})]_i \in \mathbb{R}^d$ , we can rewrite (4) as

$$Q([\nabla f(\mathbf{x})]_i) = [\nabla f(\mathbf{x})]_i - \mathbf{n}, \quad (5)$$

where  $\mathbf{n} \in \mathbb{R}^d$  is the quantization noise vector. The noise vector  $\mathbf{n}$  has only one nonzero element  $n_i$  that is bounded as  $|n_i| \leq \Delta/2$ . Hence,

$$\|\mathbf{n}\| \leq \frac{\Delta}{2}. \quad (6)$$

Throughout the paper, we use  $\mathbf{x}_k$  and  $\mathbf{x}_k^q$  to denote the  $k$ th update of  $\mathbf{x}$  before and after adding the quantization noise, respectively. An upper case letter  $S$  is used for a random variable, while a lower case letter  $s$  is used for a realization of  $S$ . We also use  $\|\mathbf{x}\|$  to denote the Euclidean norm of the vector  $\mathbf{x}$ , and we use  $Q(\cdot)$  to denote the quantization operator.

## III. QUANTIZED RANDOMIZED COORDINATE DESCENT

Here, we describe the randomized coordinate descent algorithm with quantized update. The algorithm starts from an initial point  $\mathbf{x}_0$ , and stops after a predetermined number of iterations  $T$ . Set  $\mathbf{x}_0^q = \mathbf{x}_0$ . At iteration  $(j + 1)$ , a machine  $s_{j+1} \in \{1, 2, \dots, d\}$  is randomly (with a uniform distribution) selected, who calculates  $[\nabla f(\mathbf{x}_j^q)]_{s_{j+1}}$  and then sends the quantized update  $Q([\nabla f(\mathbf{x}_j^q)]_{s_{j+1}})$ , all machines update

$$\mathbf{x}_{j+1}^q = \mathbf{x}_j^q - tdQ([\nabla f(\mathbf{x}_j^q)]_{s_{j+1}}), \quad (7)$$

where  $t$  is the step size.

**Algorithm:** Quantized Randomized Coordinate Descent

---

```

1:  $\mathbf{x}_0^q = \mathbf{x}_0$ 
2: for  $j = 0, 1, \dots, (T - 1)$  do
3: a machine is randomly selected to send its update
4: selected machine  $s_{j+1}$  computes  $[\nabla f(\mathbf{x}_j^q)]_{s_{j+1}}$ 
5: machine  $s_{j+1}$  communicates  $Q([\nabla f(\mathbf{x}_j^q)]_{s_{j+1}})$ 
6: all machines update  $\mathbf{x}_{j+1}^q = \mathbf{x}_j^q - tdQ([\nabla f(\mathbf{x}_j^q)]_{s_{j+1}})$ 
7: end for
```

---

To facilitate the analysis, we also record the sequence

$$\mathbf{x}_{j+1} = \mathbf{x}_j^q - td[\nabla f(\mathbf{x}_j^q)]_{s_{j+1}}. \quad (8)$$

Using (5), we can show that

$$\mathbf{x}_j^q = \mathbf{x}_j + td\mathbf{n}_j, \quad j = \{1, 2, \dots, T\}. \quad (9)$$

It is desirable that the algorithm converges within  $k$  iterations to an accuracy level of  $\epsilon$  and a confidence level of  $\rho \in (0, 1)$ , such that

$$\Pr(\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \epsilon) \geq 1 - \rho. \quad (10)$$

By applying Markov inequality, the convergence condition in (10) is achieved if

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \epsilon\rho. \quad (11)$$

## IV. CONVERGENCE ANALYSIS

In this section, we analyze the convergence rate of the quantized randomized coordinate descent algorithm.

*Theorem 1:* Given that the quantization error  $\Delta$  is bounded as following

$$\Delta \leq \frac{\epsilon\rho L^2}{2m} \left(\frac{1}{C_{min}} - 1\right),$$

the number of iterations required for the quantized randomized coordinate descent algorithm to converge to the optimal solution  $\mathbf{x}^*$  is at most

$$k^q = \frac{\log(2\|\mathbf{x}_0 - \mathbf{x}^*\|^2/\epsilon\rho)}{\log(1/C_{min})} + \frac{\log(2\|\mathbf{x}_0 - \mathbf{x}^*\|^2)}{\log(1/(C_{min} + \frac{\epsilon\rho}{2}(1 - C_{min})))},$$

where

$$C_{min} = 1 - \frac{1}{g^2 d}.$$

*Proof:* We have that

$$\begin{aligned} \|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_j^q - \mathbf{x}^* - td[\nabla f(\mathbf{x}_j^q)]_{s_{j+1}}\|^2 \\ &= \|\mathbf{x}_j^q - \mathbf{x}^*\|^2 + t^2 d^2 \|\nabla f(\mathbf{x}_j^q)\|_{s_{j+1}}^2 \\ &\quad - 2td\langle \nabla f(\mathbf{x}_j^q) \rangle_{s_{j+1}}, \mathbf{x}_j^q - \mathbf{x}^* \rangle. \end{aligned} \quad (12)$$

Taking the expectation of both sides with respect to the independent and identically distributed (i.i.d.) random variables  $S_1, S_2, \dots, S_{j+1}$

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 &= \mathbb{E}\|\mathbf{x}_j^q - \mathbf{x}^*\|^2 \\ &\quad + t^2 d^2 \mathbb{E}\|\nabla f(\mathbf{x}_j^q)\|_{s_{j+1}}^2 \\ &\quad - 2td\mathbb{E}\langle \nabla f(\mathbf{x}_j^q) \rangle_{s_{j+1}}, \mathbf{x}_j^q - \mathbf{x}^* \rangle. \end{aligned} \quad (13)$$

Since  $\mathbb{E}_{s_{j+1}}[\nabla f(\mathbf{x}_j^q)]_{s_{j+1}} = \frac{1}{d}(\nabla f(\mathbf{x}_j^q))$ , then

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 &= \mathbb{E}\|\mathbf{x}_j^q - \mathbf{x}^*\|^2 + t^2 d \mathbb{E}\|\nabla f(\mathbf{x}_j^q)\|^2 \\ &\quad - 2t\mathbb{E}\langle \nabla f(\mathbf{x}_j^q), \mathbf{x}_j^q - \mathbf{x}^* \rangle. \end{aligned} \quad (14)$$

By applying inequalities (2) and (3), and using the fact that  $\nabla f(\mathbf{x}^*) = 0$ , we have that

$$\|\nabla f(\mathbf{x}_j^q)\| \leq L\|\mathbf{x}_j^q - \mathbf{x}^*\|, \quad (15)$$

and

$$\langle \nabla f(\mathbf{x}_j^q), \mathbf{x}_j^q - \mathbf{x}^* \rangle \geq m\|\mathbf{x}_j^q - \mathbf{x}^*\|. \quad (16)$$

Substituting (15) and (16) in (14), we get that

$$\mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 \leq C\mathbb{E}\|\mathbf{x}_j^q - \mathbf{x}^*\|^2, \quad (17)$$

where  $C = t^2 L^2 d - 2tm + 1$ . We also have that

$$\begin{aligned} \|\mathbf{x}_j^q - \mathbf{x}^*\|^2 &= \|\mathbf{x}_j - \mathbf{x}^* + td\mathbf{n}_j\|^2 \\ &= \|\mathbf{x}_j - \mathbf{x}^*\|^2 + t^2 d^2 \|\mathbf{n}_j\|^2 \\ &\quad + 2td\langle \mathbf{x}_j - \mathbf{x}^*, \mathbf{n}_j \rangle \\ &\leq \|\mathbf{x}_j - \mathbf{x}^*\|^2 + t^2 d^2 \|\mathbf{n}_j\|^2 \\ &\quad + 2td\|\mathbf{x}_j - \mathbf{x}^*\| \|\mathbf{n}_j\| \\ &\leq \|\mathbf{x}_j - \mathbf{x}^*\|^2 + td\Delta \|\mathbf{x}_j - \mathbf{x}^*\| \\ &\quad + \frac{t^2 d^2 \Delta^2}{4}, \end{aligned} \quad (18)$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from (6).

To proceed with the convergence analysis, we have two different cases.

**Case 1** ( $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq 1$ ):

In this case,  $\mathbb{E}\|\mathbf{x}_j - \mathbf{x}^*\| \leq 1$ . Therefore,

$$\mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 \leq C\mathbb{E}\|\mathbf{x}_j - \mathbf{x}^*\|^2 + Ctd\Delta(1 + \frac{td\Delta}{4}). \quad (19)$$

Let  $k_1$  denotes the minimum number of iterations required to achieve the convergence condition. Hence,

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{k_1} - \mathbf{x}^*\|^2 &\leq C^{k_1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\ &\quad + Ctd\Delta(1 + \frac{td\Delta}{4})(1 + C + \dots + C^{k_1-1}). \end{aligned} \quad (20)$$

Since  $C < 1$ , then

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{k_1} - \mathbf{x}^*\|^2 &\leq C^{k_1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\ &\quad + \frac{C}{1-C} t d \Delta (1 + \frac{td\Delta}{4}). \end{aligned} \quad (21)$$

For the algorithm to converge, let

$$C^{k_1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{\epsilon \rho}{2}, \quad (22)$$

and

$$\frac{C}{1-C} t d \Delta (1 + \frac{td\Delta}{4}) \leq \frac{\epsilon \rho}{2}, \quad (23)$$

**Case 2** ( $\|\mathbf{x}_0 - \mathbf{x}^*\| > 1$ ):

Let  $k_2$  denotes the minimum number of iterations required such that  $\mathbb{E}\|\mathbf{x}_{k_2} - \mathbf{x}^*\| \leq 1$ . For all  $j \leq k_2$ , we have that  $\mathbb{E}\|\mathbf{x}_j - \mathbf{x}^*\| \leq \mathbb{E}\|\mathbf{x}_j - \mathbf{x}^*\|^2$ . Therefore,

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 &\leq C(1 + td\Delta) \mathbb{E}\|\mathbf{x}_j - \mathbf{x}^*\|^2 \\ &\quad + \frac{Ct^2 d^2 \Delta^2}{4}. \end{aligned} \quad (24)$$

After  $k_2$  iterations, we have that

$$\begin{aligned} \mathbb{E}\|\mathbf{x}_{k_2} - \mathbf{x}^*\|^2 &\leq (C(1 + td\Delta))^{k_2} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\ &\quad + \frac{Ct^2 d^2 \Delta^2}{4(1-C)}. \end{aligned} \quad (25)$$

For the algorithm to converge, let

$$(C(1 + td\Delta))^{k_2} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{1}{2}, \quad (26)$$

and

$$\frac{Ct^2 d^2 \Delta^2}{4(1-C)} \leq \frac{1}{2}. \quad (27)$$

Finally, the total number of iterations required for convergence is given by

$$k^q = k_1 + k_2. \quad (28)$$

To achieve the fastest convergence rate, the step size  $t$  is chosen to minimize  $C$ . Hence,

$$t_{opt} = \frac{1}{gLd}, \text{ and } C_{min} = 1 - \frac{1}{g^2 d} \quad (29)$$

From (23) and (27), a sufficient condition on the quantization error is given by

$$\Delta \leq \frac{\epsilon \rho L^2}{2m} \left( \frac{1}{C_{min}} - 1 \right). \quad (30)$$

From (22), (26), and (30), the number of iterations required for the algorithm to converge is at most

$$\begin{aligned} k^q &= \frac{\log(2\|\mathbf{x}_0 - \mathbf{x}^*\|^2 / \epsilon \rho)}{\log(1/C_{min})} \\ &\quad + \frac{\log(2\|\mathbf{x}_0 - \mathbf{x}^*\|^2)}{\log(1/(C_{min} + \frac{\epsilon \rho}{2}(1 - C_{min})))}. \end{aligned} \quad (31)$$

Note that By setting  $\Delta = 0$  and hence  $x_j^q = x_j$  in (17), the quantization-free scenario can be recovered. It follows

that the number of iterations required for the quantization-free algorithm to converge is at most

$$k = \frac{\log(\|\mathbf{x}_0 - \mathbf{x}^*\|^2/\epsilon\rho)}{\log(1/C_{min})}, \quad (32)$$

which coincides with the result obtained in [10].

## V. SIMULATION RESULTS

In this section, we run an experiment to verify that the quantization error does not propagate and hence the convergence is possible. For that purpose, we apply the quantized randomized coordinate descent algorithm to solve a linear regression problem. The data set we use is collected from a power plant [11]. It has four predictors (Temperature, Pressure, Humidity, and Exhaust Vacuum) and one output (Electrical Energy). All data is normalized to have zero mean and a standard deviation of one. The number of observations is  $n = 9568$ .

To solve this problem, it is required to minimize the square loss function

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^n (y_i - A_{i:}\mathbf{x})^2, \quad (33)$$

where  $A$  is the data matrix,  $A_{i:}$  is the  $i$ th row of  $A$ , and  $\mathbf{y}$  is the output vector. Notice that the first column of  $A$  is a vector of ones, which is added to evaluate for the intercept. The network consists of five nodes in addition to the fusion center; the first node calculates the derivative in the direction of the intercept coefficient, while each of the remaining nodes calculates the derivative in the direction of one predictor coefficient. The algorithm starts from  $\mathbf{x}_0 = [1, 1, 1, 1, 1]^T$  and iterates to reduce the coefficients residual  $\|\mathbf{x}_j - \mathbf{x}^*\|^2$ .

**Experiment 1:**  $t = 10^{-4}$ ,  $\Delta = 10^5$ .

First, we plot the coefficients residual against the number of iterations as shown in Fig. 1.

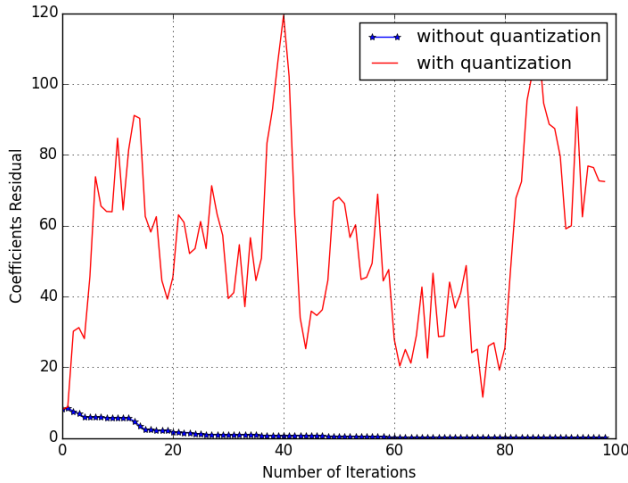


Fig. 1: Effect of the quantization error ( $\Delta = 10^5$ ) on the coefficients residual.

Then, we plot the predicted value for an input of all ones  $A_1 = [1, 1, 1, 1, 1]$  against the number of iterations as shown in Fig. 2.

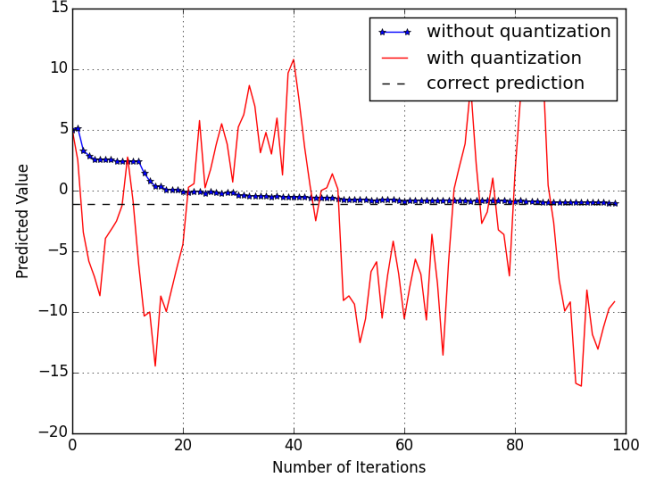


Fig. 2: Effect of the quantization error ( $\Delta = 10^5$ ) on the predicted value.

Figures 1 and 2 show that the quantized randomized coordinate descent algorithm diverges if the quantization error  $\Delta = 10^5$ . This result is intuitive since a large quantization error is expected to prevent the algorithm from converging to the optimal solution.

**Experiment 2:**  $t = 10^{-4}$ ,  $\Delta = 10^3$ .

First, we plot the coefficients residual against the number of iterations as shown in Fig. 3.

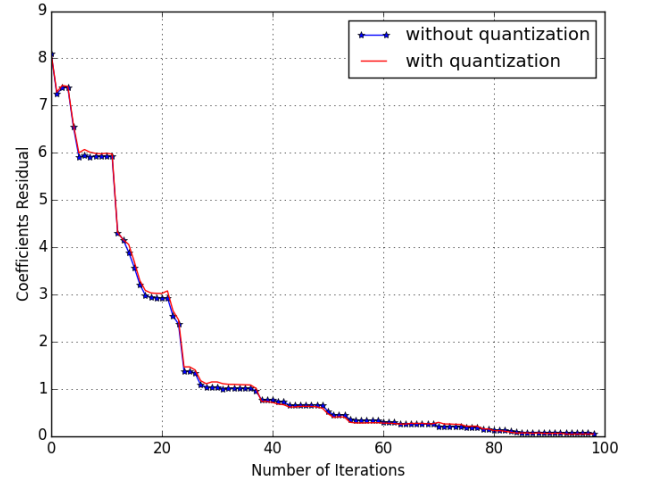


Fig. 3: Effect of the quantization error ( $\Delta = 10^3$ ) on the coefficients residual.

Then, we plot the predicted value for an input of all ones  $A_1 = [1, 1, 1, 1, 1]$  against the number of iterations as shown in Fig. 4.

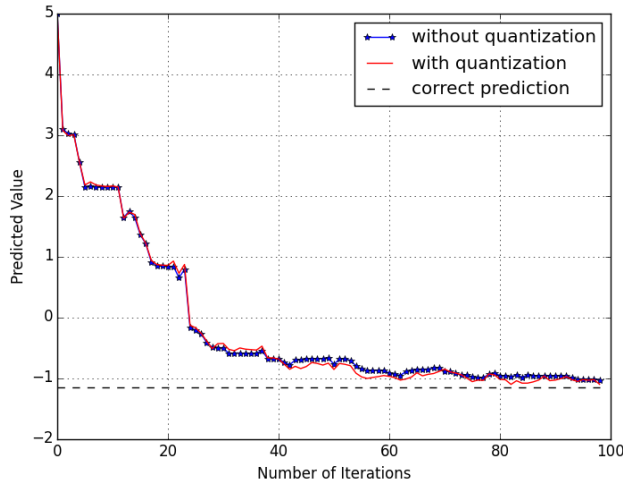


Fig. 4: Effect of the quantization error ( $\Delta = 10^3$ ) on the predicted value.

Figures 3 and 4 show that the quantized randomized coordinate descent algorithm converges for a smaller value of the quantization error  $\Delta = 10^3$ . This verifies that the quantization error does not propagate and hence the convergence is possible if the quantization error is bounded, which coincides with the result we obtained in Theorem 1.

## VI. CONCLUSION

In this paper, we have studied the problem of distributed optimization under communication constraints. We have modified the randomized coordinate descent algorithm to solve an unconstrained convex minimization problem in the presence of quantization error. We have analyzed the convergence rate of our algorithm, and we have derived sufficient conditions on the quantization error to guarantee that the algorithm converges to the optimal solution. We have further verified that the convergence is possible in the presence of quantization error by running an experiment that solves a linear regression problem.

## REFERENCES

- [1] M. Jaggi, V. Smith, M. Takáč, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan, *Coordinate descent algorithms*. In *Advances in NIPS* 27, pp. 3068–3076, 2014.
- [2] Y. Zhang and L. Xiao, *DiSCO: Distributed optimization for self-concordant empirical loss*. In *ICML*, pp. 362–370, 2015.
- [3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2010.
- [4] W. Deng and W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*. *Journal of Scientific Computing*, pp. 1–28, 2012.
- [5] S. Zhu, M. Hong, and B. Chen, “Quantized consensus admm for multi-agent distributed optimization,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Shanghai, China), pp. 4134–4138, Mar. 2016.
- [6] S. J. Wright, *Coordinate descent algorithms*. *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

- [7] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*. *Mathematical Programming*, pages 1–38, 2012.
- [8] R. Tappenden, P. Richtárik, and J. Gondzio, “Inexact coordinate descent: complexity and preconditioning,” *Journal of Optimization Theory and Applications*, vol. 170, pp. 144–176, Jul. 2016.
- [9] P. Richtárik and M. Takáč, *Distributed coordinate descent method for learning with big data*. arXiv:1310.2059, 2013.
- [10] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, *Perturbed iterate analysis for asynchronous stochastic optimization*. arXiv:1507.06970v2, 2016.
- [11] P. Tufekci, *Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods*. *International Journal of Electrical Power and Energy Systems*, vol. 60, pp. 126–140, 2014.